The effect of non-labial facial information on audiovisual speech perception

Zeyu Huang[1, a], Yao Lu[1, b], Lu Wang[1, c] and Xiyu Wu[2, d]

[1]Department of Chinese Language and Literature, Peking University, Beijing 100871, China

[2] Center for Chinese Linguistics, Peking University, Beijing 100871, China

[a]zeyu.huang@pku.edu.cn, [b] luyiru2012@163.com, [c] wanglunj@163.com, [d] xiyuwu@pku.edu.cn

Keywords: McGurk effect, audiovisual speech perception, non-labial facial information.

Abstract. We conducted an experiment consisting of five blocks to examine the effect of non-labial facial information on audiovisual speech perception. 20 Chinese native speakers were asked to report the syllables they perceived during five condition: audio-only, video-only, video-only without labial part, audiovisual and audiovisual without labial part. The materials were /pa/, /tsa/, /ta/, /tʂa/ and /ka/, which were selected according to places of articulation from front to back. The results showed that even though the non-labial facial information was not enough to distinguish non-labial consonants, they could have significant effect on auditory speech perception.

1. Introduction

Although speech perception in natural condition was a multisensory process, classic models of speech processing focused predominantly on acoustic input, ignoring the influence of visual information (Van Wassenhove V, 2013). As a meter of fact, visual input does not only provide subsidiary information such as identification or emotion, the forms and kinematics of facial information could also provide abundant details of articulation, which could even affect the speech processing directly and cause a fused illusion when video and audio input were incongruent (McGurk H, MacDonald J, 1976).

However, it was unclear how we extract articulation information from visual input and what parts of visual information work during speech processing. Several studies have indicated that mouth was not the only resource for perceiving linguistic information (Rosenblum L D, Saldaña H M, 1996; Paré, et al., 2003). Even if the fixation point was fixed 10°-20° from talker's mouth, McGurk effect persisted. Therefore, the present study was to explore whether and to what extent non-labial facial information could affect visual and audiovisual speech perception.

2. Method

2.1 Subjects

20 Mandarin speakers including 8 males and 12 females ranging from 19 to 29 years old (overall mean age=23.6±2.53 years) attended this research. All of them had normal or corrected-to-normal vision and no speech or hearing impairment. None of them had received lip reading training. They had no idea of experiment hypothesis all the way.

2.2 Stimuli

The audiovisual stimuli for the experiment were recorded by a EOS kiss X5 camera and a professional external microphone in the studio of Linguistic Laboratory of Peking University. The frame rate of video was 29.97 FPS and the sampling rate of audio was 48kHz.

The stimuli were made of 2 native speakers of Mandarin, 1 male (m1) and 1 female (f1). Only the head and shoulder were shot against a dark blue background. The materials were edited with Adobe Premiere 2018 to ensure that each stimulus was 2-second long and without blinks.

There were 5 Chinese syllables /pa/, /tsa/, /ta/, /tʂa/ and /ka/ which were composed of vowel /a/ and a series of consonants, each represented for a place of articulation from front to back. For incongruent audiovisual stimuli, /pa/ was dubbed into the videos of the other syllables. Because according to previous studies, the McGurk effect arose by audio /pa/ tended to be the strongest among all kinds of incongruent pairs of Chinese syllables (Pan X, 2011). Besides, for the series of non-labial stimuli, the mouth areas were covered by black oval masks, which were set by the frame of each stimulus when mouth was open widest. To sum up, there were 66 stimuli for all, including 10 (2*5) stimuli for each of audio-only (OA), video-only (OV), video-only without mouth area (OV_NoM) condition, and 18 stimuli (10 congruent and 8 incongruent) for each of audiovisual (AV) and audiovisual without mouth area (AV_NoM) condition. For each block, the stimuli were presented randomly.

3. Results

3.1 Audio Only

When there were only audio stimuli available, the recognition rates were as followed:

Table 1 Percentages of Correct Identifications of Audio-Only (OA) Condition

	STIMULI					
Talker	/pa/	/tsa/	/ta/	/tʂa/	/ka/	Average
f1	100.00	100.00	100.00	95.45	81.82	95.454
m1	81.82	100.00	100.00	95.45	86.36	92.726

The average percentage of all audio stimuli was 93.81%±6.69%. According to a two-way ANOVA, the main effect of talker was not significant [F(1, 20)=1.00, p=0.329)]. The identification rate of /ka/ was significantly smaller than /tsa/ and /ta/. All audio stimuli could be identified at high proportion.

3.2 Video Only

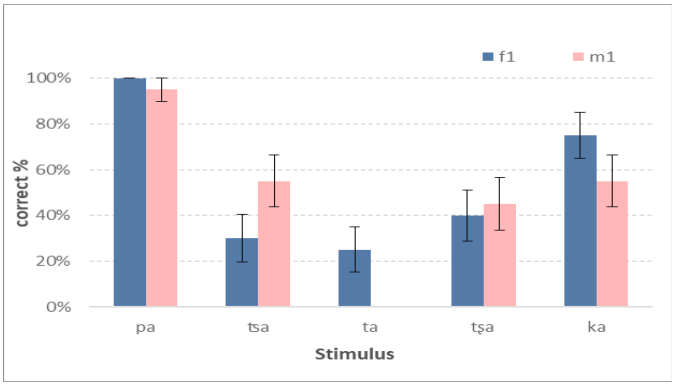

Fig. 1 Percentages of Correct Identifications of Video-Only (OV) Condition

Figure 1 shows the rate of correct identifications when only visual stimuli were presented. Analyzed with a two-way ANOVA, there was a significant interaction effect between talker and stimulus [F(4,76)=3.673, p<0.01)]. We conduct a paired comparison adjusted by Bonferroni test, for /ta/ and /ka/, the identification rate of f1 was significantly higher than those of m1. And for each talker, the accuracy of labial consonant /p/ was much higher than the other non-labial consonants,

while the /t/ sound was the lowest. Overall, the percentages of correct identifications distributed in U-shape according to places of articulation, which was considerable high for labial consonant, then decrease sharply to bottom, and rise again at the place of velar.

3.3 Video Only without Mouth Area

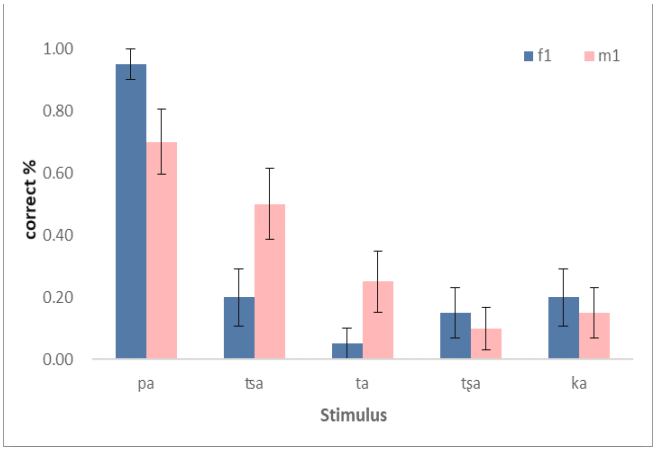

Fig. 2 Percentages of Correct Identifications of Video-Only without mouse area (OV_NoM) Condition

For the block of visual stimuli without mouth, compared with normal visual condition, the accuracy of all stimuli decrease to some extent. There was a significant interaction effect between talker and stimulus [$F(4,76)=3.956$, $p<0.01$)]. For /pa/, accuracy of f1 was significantly higher than m1. There was no significant difference between f1 and m1 on other syllables. What's more, for each talker, the percentages of correct identification of these consonants were decline with places of articulation from front to back. Compared to OV condition, the identification rate of /ka/ was affected most, then was /tʂa/. The other syllables were slightly or barely influenced.

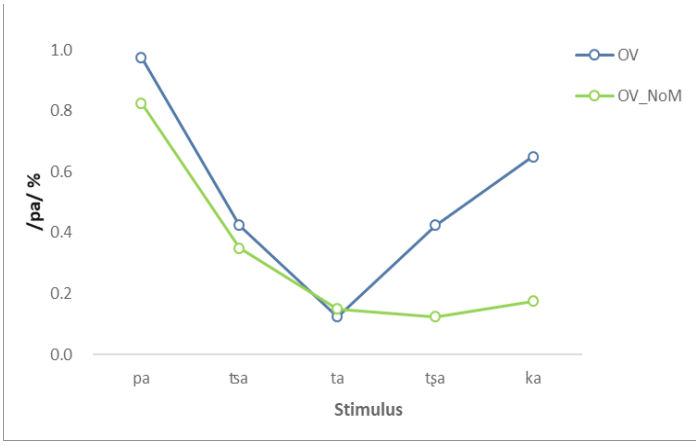

Fig. 3 Percentages of Correct Identifications of OV(Video-Only) versus OV_NoM (Video-Only without mouse area) condition

3.4 Audiovisual

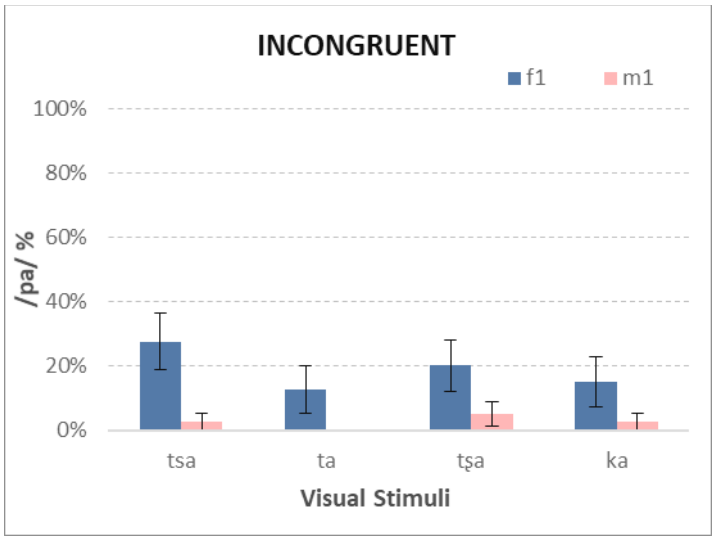

Fig. 4 Accuracy performance in response to the presentation of incongruent audiovisual stimuli (AV).

The percentages of correct identifications of audiovisual condition were showed in Figure 4. No interaction effect was significant between talker and visual stimuli[$F(3,16)=1.380$, $p=0.285$)]. And visual stimulus has no main effect on McGurk effect [$F(3,16)=3.198$, $p=0.052$)]. However there was a significant different between two talkers[$F(1,18)=6.133$, $p<0.05$)]: the accuracy percentage of f1 (19.7±7.2%) was much higher than m1(2.6%±1.5%), which means the McGurk effect of f1 was much weaker than m1, though by and large, the McGurk effect of every stimulus was considerably strong.

3.5 Audiovisual without Mouth Area

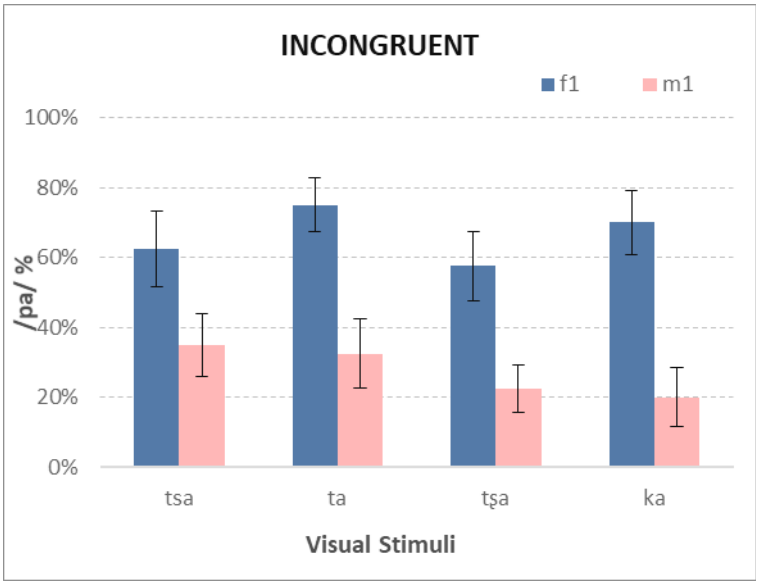

Fig. 5 Accuracy performance in response to the presentation of incongruent audiovisual stimuli without mouth (AV_NoM).

When mouth areas were covered, there was no significant interaction between talker and visual stimuli[$F(3,57)=1.132$, $p=0.344$)], and no significant difference among visual stimuli [$F(3,57)=1.830$, $p=0.152$)]. However the main effect of talker was still significant [$F(1,19)=24.057$, $p<0.01$)]: the accuracy percentage of f1 (66.3±7.4%) was much higher than m1(27.5%±6.9%), that was to say the McGurk effect of f1 was still weaker than m1.

Compared with AV condition, no other effect or interactions among talker, mouth condition and visual stimuli were found to be statistically significant, except for the interaction between talker and mouth condition[$F(1,18)=12.906$, $p<0.01$)]. Even though the McGurk effect of all stimuli tend to be weaker when mouth areas were covered, there is still 1/3 to 1/2 chance that fusion illusions occurred.

4. Conclusion

The results showed that McGurk effect never disappeared even when mouth areas were totally covered, which indicated that even though the non-labial facial information was not enough to identify non-labial consonants by itself, the visible kinematics of articulatory gestures on the non-labial facial area could have significant effect on auditory speech perception.

References

[1] Van Wassenhove V. Speech through ears and eyes: interfacing the senses with the supramodal brain[J]. Frontiers in Psychology, 2013, 4(JUL): 1–17.

[2] Sumby W H, Pollack I. Visual Contribution to Speech Intelligibility in Noise[J]. The Journal of the Acoustical Society of America, 1954, 26(2): 212–215.

[3] Mcgurk H, Macdonald J. Hearing lips and seeing voices[J]. Nature, 1976, 264(5588): 746–748.

[4] Green K P, Kuhl P K, Meltzoff A N et al. Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect[J]. Perception & Psychophysics, 1991, 50(6): 524–536.

[5] Munhall K G, Gribble P, Sacco L et al. Temporal constraints on the effect.[J]. Perception & Psychophysics, 1996, 48(3): 351–362.

[6] Sekiyama K. Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subjects[R]. 1997, 59(1).

[7] Sekiyama K, Tohkura Y. McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility[J]. The Journal of the Acoustical Society of America, 1991, 90(4): 1797–1805.

[8] Macdonald J, Andersen S, Bachmann T. Hearing By Eye: Visual Spatial Degradation And The Mcgurk Effect[C]//Sixth European Conference on Speech Communication and Technology. 1999.

[9] Rosenblum L D, Saldaña H M. An audiovisual test of kinematic primitives for visual speech perception.[J]. Journal of Experimental Psychology: Human Perception and Performance, 1996, 22(2): 318–331.

[10] Munhall K G. Eye movement of perceivers during audiovisual speech perception[J]. Perception & Psychophysics, 1998, 60(6): 926–940.

[11] Martin Paré, Rebecca C. Richler M T H G M. Gaze behavior in audiovisual speech perception: The influence of ocular fixations on the McGurk effect[R]. 2003.

[12] Worster E, Pimperton H, Ralph-Lewis A Et Al.. Eye Movements During Visual Speech Perception in Deaf and Hearing Children[J]. Language Learning, 2017(June): 159–179.

[13] Pan, X. Labial Coaticulation and Audio-Visual Speech Perception in Standard Chinese [M]. 2011.